Final Report

Nonlinear Auditory Modeling as a Basis for Speaker Recognition

T.F. Quatieri

17 May 2002

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

ADA402327

Massachusetts Institute of Technology
Lincoln Laboratory

Nonlinear Auditory Modeling as a Basis for Speaker Recognition

T.F. QUATIERI
Group 62

FINAL REPORT

17 May 2002

Lexington   Massachusetts

ABSTRACT

In this report, we develop a front-end nonlinear auditory model based on recent work of Dau, Puschel, and Kohlrausch (DPK) [Dau, Puschel, and Kohlrausch, 1997]. An important aspect of the model is the robust accentuation of temporal change in a signal at the cochlea level that forms the basis of a feature set for automatic speaker recognition. Preliminary speaker recognition experiments with the DPK front-end auditory model give performance close to that from the standard mel-cepstrum. Fusion of scores from the mel-cepstrum and the DPK front-end auditory model, however, is shown to give a useful performance gain relative to the standard mel-cepstrum alone. The dynamics provided by the nonlinear auditory model, therefore, appears to provide some "orthogonality" to that of the more static mel-cepstral representation.

In addition, in this report, we provide initial development of new "common modulation" features based on modeling a more central region of auditory processing in the brain's *inferior colliculus* than the low-level auditory front-end. These higher-level features rely on the DPK auditory model as a foundation for further analysis of low-level temporal trajectories. This new feature representation is an important research direction and provides additional feature "orthogonality" to front-end auditory processing, as exhibited in improved speaker recognition performance with fusion of scores from low-level and high-level feature sets.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Continued)

1.0 INTRODUCTION

The proposed primary objective for FY2001 was to develop robust techniques for extraction of speech source and vocal tract parameters and to apply these techniques to improve speaker recognition in degrading environments characterized by noise and distortion. For the robust speaker recognition research, our approach is to develop new parameter extraction techniques using linear and nonlinear models of distributed source generation in the vocal tract, and more generally using novel temporal representations of speech events on different time scales. In prior work at Lincoln [Quatieri, Jankowski, and Reynolds 1994], speaker identification performance improvements have been achieved by estimating the onset times of secondary excitation pulses within glottal cycles. Here we had   assumed secondary excitations (per glottal cycle) were associated with a nonlinear production model, e.g., multiple vocal fold vibrations or sound generation by vortices. This is a special case of distributed source generation illustrated in Figure 1.

Figure 1: Distributed excitation modeling approach.
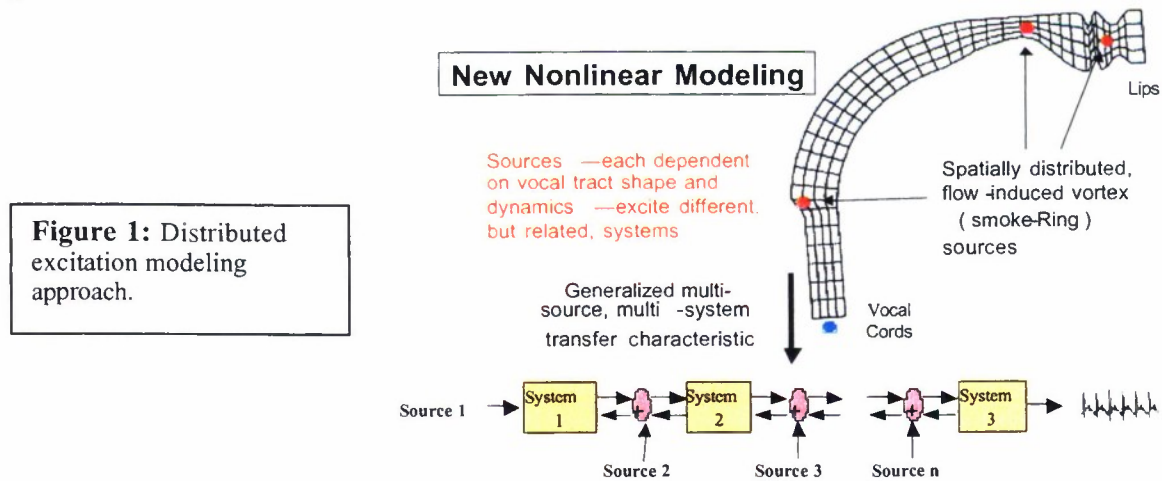
In onset-time recognition experiments, the source *arrival times* were measured with the high-resolution Teager energy operator that consists of only three discrete-time samples [Maragos, Kaiser, Quatieri, 1996] and gives a robust high-resolution energy estimate. The onset times were used as features in the Lincoln Gaussian-mixture speaker identification algorithm [Reynolds, 1995]. Feature sets were constructed with primary pitch and secondary pulse locations derived from low and high speech formants (Figure 2). Preliminary testing was performed with a confusable 40-speaker subset from the NTIMIT (telephone channel) database. Speaker identification improved from 55% to 70% correct classification when the full set of new resonant energy-based features were added as an independent stream to conventional mel-cepstra [Quatieri, Jankowski, Reynolds, 1994]. About half of the performance gain was achieved with the secondary pulse arrival times. One important observation in this work is that occasionally within a glottal cycle, some formants exhibited both primary and secondary onset times, while others only primary onset times, and that use of the secondary onset times as a speech feature improved speaker identification performance. This is consistent with secondary pulses not always exciting all formants associated with primary pulses, and may be consistent with a speaker dependence of this attribute.

ORIGINAL WAVEFORM

TEAGERY ENERGY WITH PRIMARY AND SECONDARY PULSE EXAMPLES

0.000   0.005   0.010   0.015   0.020   0.025   0.030   0.035   0.040

TIME (SECONDS)

Figure 2: Estimation of distributed source time-of-arrivals from the Teager energy formant output.
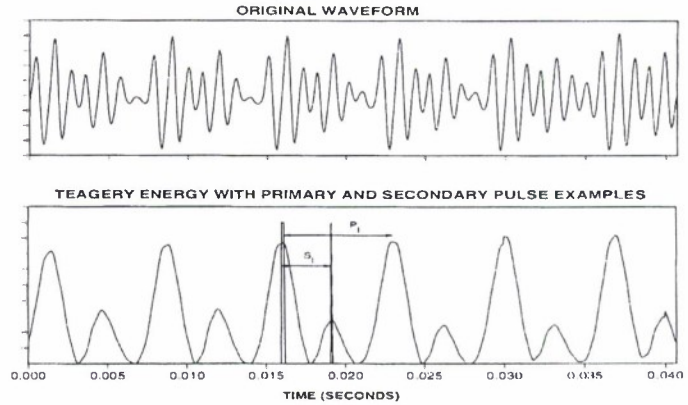
The importance of these experiments for our purpose is the illustration of the existence of *temporal-based features* "orthogonal" to the spectral-based mel-cepstrum. Nevertheless, there are limitations to this approach in requiring explicit formant, pitch, voicing, and pulse-time estimation that are not desirable under degrading conditions. Furthermore, the approach is far too model-constrained to reflect a rich variety of temporal features. A more practical problem is that a C-code rendition of the current Matlab code is a large time investment. Although we may revisit this approach in the future, we now take a different tact. We desire this different approach to give a rich variety of temporal features that reflect both linear (e.g., transient and transitional) and nonlinear (e.g., distributed sources from vortices or secondary glottal flaps) phenomenon, allow perspectives on different time scales (e.g. within a single glottal cycle and a phone sequence), and be robust in degrading environments characterized by additive noise and signal distortions.

With these desired properties in mind, we have introduced an auditory model-based approach because of its known ability to represent temporal structure on both a fine and coarse scale [Pickles, 1988], as from complex nonlinear production models, and because of its speech recognition robustness in noise for human and machine [Strobe and Albeer, 1998], [Tchorz and Kollmeier, 1999], [Schmidt-Nielsen and Crystal, 2000]. This perception-based, non-parametric approach is a more solid foundation than the parametric approach that originally motivated this study in which secondary pulse times provided temporal feature streams [Quatieri, Jankowski, and Reynolds, 1994] because, as we have observed, the later approach requires explicit formant, pitch, and voicing estimation that is not desirable under conditions. Specifically, we are utilizing a recent auditory perception model by Dau, Puschel, and Kohlrausch that accounts for importance of temporal change arising from speech dynamics [DPK, 1996]. We are also using a more-recent extension of this model to a higher level of the auditory pathway, motivated by the physiological studies of Schreiner and Langner [Schreiner and Langner, 1988] that purport that temporal amplitude modulations within the *inferior colliculus* of the cat are *"orthogonal" to the more spectrally-based lower-level front-end auditory features.*

In this report we show, in the context of automatic speaker recognition, some degree of "orthogonality" with the mel-cepstrum of two new auditory-based feature sets: (1) The nonlinearity of the front-end auditory model provides accentuated dynamics not seen by a conventional front-end auditory model and (2) A high-level (inferior colliculus) frequency analysis and integration of common modulation of the auditory channels of (1) further reveals temporal structure. Fusion of likelihood scores from these two feature sets with mel-cepstral scores provides an encouraging recognition performance gain over use of the conventional mel-cepstrum alone. Specifically, this report describes the following FY01 accomplishments:

2

1. C-code implemention of the Dau, Puschel, and Kohlrausch (DPK) front-end nonlinear auditory model [Dau, Puschel, and Kohlrausch, 1997] that robustly accentuates temporal change.

2. Preliminary automatic speaker recognition experiments with the DPK front-end auditory model, giving speaker recognition performance close to that from the standard mel-cepstrum.

3. Fusion of scores from the mel-cepstrum and the DPK front-end auditory model, giving useful performance gain relative to the mel-cepstrum alone. The dynamics provided by the full auditory model, therefore, appears to provide some "orthogonality" to that of more static features.

4. Initial development of new "common modulation" features based on a more central region of auditory processing in the brain's inferior colliculus than the low-level auditory front-end. These higher-level features rely on the DPK auditory model as a foundation for further analysis of low-level temporal trajectories. This new feature representation is an important research direction and provides additional feature "orthogonality" to front-end auditory processing and, as a result, improved speaker recognition performance.

In this report, we summarize our accomplishments in each of the above areas, and end with a discussion of future directions.

2.0 FRONT-END AUDITORY MODEL DEVELOPMENT

We observed above that the front-end auditory model of "effective signal processing" by Dau, Puschel, and Kohlrausch (DPK) accounts for importance of temporal change (i.e., speech dynamics) [DPK, 1996] and gives improved automatic speech recognition relative to mel-cepstrum in noise [Strope and Albeer, 1997], [Tchorz and Kollmeier, 1999]. In this section, we describe our simulation of the DPK model, modification of the model for specific desired temporal resolution, and examples using both synthetic and real speech. We begin with a baseline system that gives equivalent temporal resolution to the standard mel-cepstrum and does not invoke the nonlinear component of the model. The nonlinear element, responsible for auditory adaptation, is then added, resulting in enhanced manifestation of temporal dynamics.

2.1 Baseline System

The three basic elements of the DPK front-end model are: (1) Constant-Q filter-bank front end, (2) Adaptive nonlinear compression, and (3) Smoothing of nonlinearly-compressed channel outputs (Figure 3). In the original DPK model, the smoothing component was fixed across frequency to emulate certain aspects of temporal masking; as seen in Figure 3, we have generalized the smoother to vary with auditory channel. In this section, we describe the two linear modules and in the following section, we describe the nonlinear module.

Figure 3: Basic elements of the Dau, Puschel, and Kohlrausch auditory model.

Cochlear — Auditory-Like Filter Bank $h_k[n]$

Adaptation — Nonlinearity

Post-smoothing — Band-Dependent Smoothing $g_k[n]$
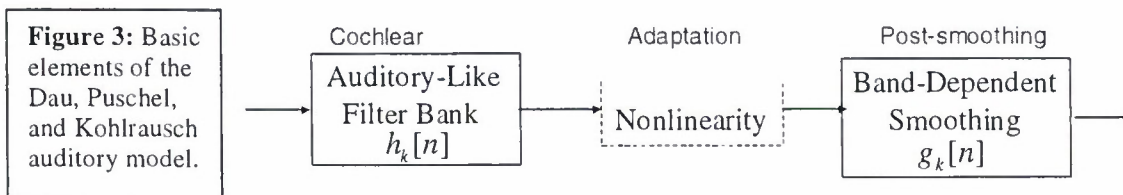
The bandpass filters of the auditory-like filter bank have spectral magnitude envelopes that are Gaussian in shape. The bandwidths and center frequencies of each Gaussian filter were designed to be close to that of triangular filters used in the standard mel-filter energy calculation. These Gaussian filters are illustrated in Figure 4 superimposed on the triangular filters used by the mel-cepstrum. This particular design was chosen in order to avoid the discontinuous derivative of the triangular filters while providing a basis similar to the triangular filters. To implement the Gaussian filter bank, zero phase was assigned to each Gaussian-shaped envelope and inversed transformed to form a set of FIR filters, each 240 samples (30 ms at 8000 samples/s) in duration. Each filter is convolved with the pre-emphasized input speech waveform via FFT overlap-add. Pre-emphasis was also used in the original DPK model; we selected the first-order pre-emphasis that is used in the standard mel-cepstrum calculation in order to provide a baseline as close as possible to the mel-cepstrum formulation.
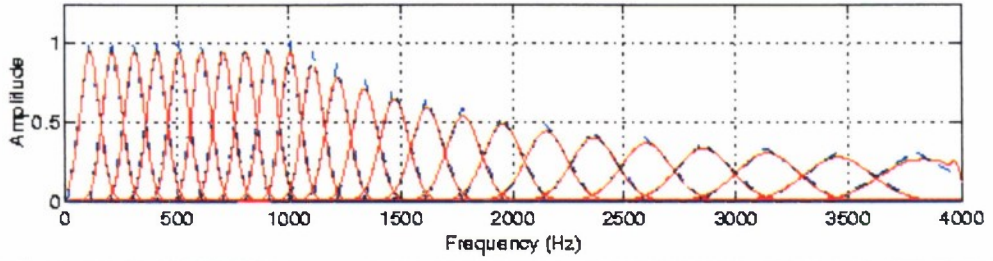

Figure 4: The band-pass filters used in the DPK front end are uniformly spaced up to 1000 Hz and have logarithmic spacing and bandwidth above 1000 Hz up to 4000 Hz. The shape of each filter frequency response is Gaussian (solid curves) and selected to match each triangular shape from the standard mel-filters (dashed curves). To preserve phase, we chose a zero-phase rendition of the Gaussian-shaped filters. Further improvements might be made with replacing the Gaussian filters with Gammatone filters (as used by DPK) and by forming a minimum-phase rendition of the filter responses, both of which are known to better emulate auditory signal processing.

A band-dependent smoother was selected to yield resolution approximately equivalent to a 20-ms Hamming window when convolved with each filter-output envelope of the filter bank. The resulting resolution is thus equivalent to that of the standard mel-cepstrum derived from a 20-ms Hamming window analysis. Specifically, for the kth Gaussian filter $h_k[n]$ of our filter bank, and a Hamming window $w[n]$, the output smoothing filter $g_k[n]$ is designed such that $w[n] = g_k[n] * h_k[n]$ where * denotes convolution and where $h_k[n]$ is understood to be complex. Each filter $g_k[n]$ was obtained using an inverse FFT of $W(\omega)/H_k(\omega)$ where $W(\omega)$ and $H_k(\omega)$ denote the frequency response of $w[n]$ and $h_k[n]$, respectively. We refer to $g_k[n]$ as the "exact inverse" smoothing filters. Each resulting filter was constrained to be 30 ms in duration by truncation.

An example of an exact inverse design is shown in Figure 5 for the 23rd channel of a 24-channel Gaussian filter bank. The bandwidth of the 23rd channel is very wide, and thus its impulse response very narrow, and so for this case a long exact inverse smoothing filter is required to meet the 20-ms resolution of the Hamming window. For lower-frequency filters, shorter filter

4

responses are required to meet the desired time resolution. Using features derived by sampling the resulting filter-output envelopes, we found speaker recognition performance, using the corpus and evaluation strategy of the following section, to be roughly equivalent to that of the standard mel-cepstral analysis, as expected because we have not attempted to improve on the window resolution. By forcing a temporal resolution equivalent to that of the mel-cepstrum, we have removed any possible recognition benefit of the greater inherent temporal resolution provided by a direct auditory filter-like output, a path we will revisit in the future. Nevertheless, this analysis system provides a good baseline for establishing the importance of auditory-model adaptation and higher-level representations.

Figure 5: Design of an "exact inverse" smoothing filter for channel 23 of a 24-channel Gaussian filter bank. Upper panel: Envelope of Gaussian filter impulse response. Middle panel: Exact inverse design. Lower panel: Superposition of 20-ms Hamming window (solid) and convolution of Gaussian filter and exact inverse (dashed) filter impulse responses.
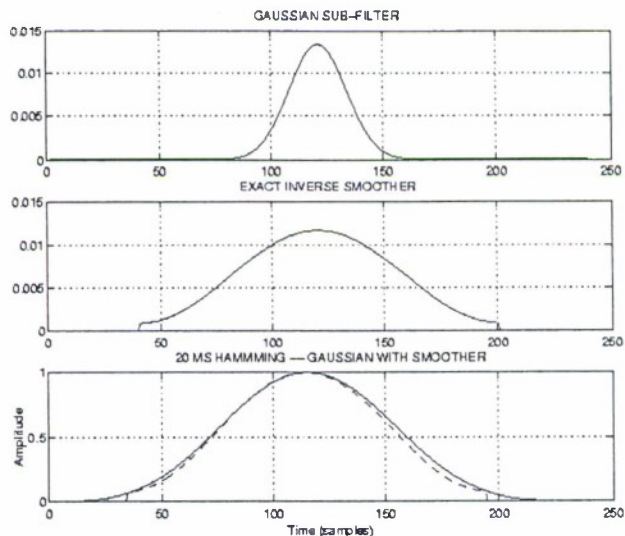
2.2 Nonlinear Adaptation

The adaptive nonlinear component of the DPK model enhances temporal change and appears to be the most important element of the model in achieving robustness in noise [Tchorz and Kollmeier, 1999]. As illustrated in Figure 6, the nonlinearity consists of five stages, each of which is a first-order smoother whose input is the output of the previous stage divided by the output of the first-order smoother delayed by one time sample. Observe that, because a division takes place at the input of each stage, one must take care in division by zero. To avoid division by exceedingly small numbers, the input level is set to a minimum threshold if it is smaller than the threshold. In our implementation, the threshold was set to 10^{-5} (over a signal range [-1,1]). Thus, for a constant input signal with zero amplitude, the output signal is not zero but the 2^5th root of 10^{-5} (with 5 adaptation loops). [In steady-state, the output O of each recursive stage is related to its input I by $O = I / O$ or $O = I^{-2}$.] Since we assume zero initial condition, we set the initial condition in each first-order smoother according to this relation for an assumed input 10^{-5}. Thus the initial condition for the first smoother is 10^{-5}, for the second 10^{-10}, and so on up to the fifth stage. After direct feedback from the authors (DPK), we were able to accurately mimic their published test cases [DPK, 1996], incorporating the above (unpublished) parameter specifications.
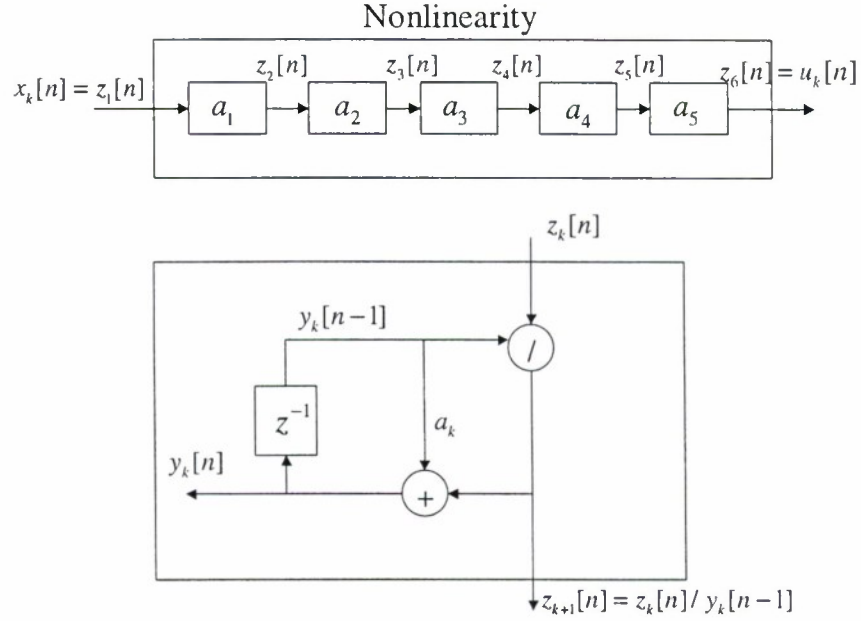
5

Nonlinearity

$x_k[n] = z_1[n]$ $z_2[n]$ $z_3[n]$ $z_4[n]$ $z_5[n]$ $z_6[n] = u_k[n]$

a_1 a_2 a_3 a_4 a_5

$z_k[n]$

$y_k[n-1]$

z^{-1}

a_k

$y_k[n]$

$z_{k+1}[n] = z_k[n] / y_k[n-1]$

Figure 6: Nonlinear module of the DPK auditory model. Upper panel: Complete five-stage nonlinearity; Lower panel: One nonlinear smoothing component. $x_k[n]$ and $u_k[n]$ are the input and ouput, respectively, of the nonlinearity of each channel.

The implementation of the full DPK auditory model, including the above nonlinearity, was first coded in Matlab for development and refinemnent purposes. Its C-code rendition was then implemented for computational speed in recognition experiments. An important difference between the Matlab- and C-code versions is the need to structure the C code frame-by-frame, while the Matlab function runs on the entire speech waveform. The C-code version thus entailed working out the buffering, level, and initial condition requirements for frame-based processing, and giving energy trajectories similar to those of the Matlab simulation. Therefore, significant effort was made in refining considerations on each component of the 5-stage nonlinearity for a frame-based system, including the limiting conditions for small-valued input.

An example in Figure 7 shows the response of the C-code system for a sine-wave input of 875 Hz, corresponding to the center frequency of the 9th auditory bandpass filter. The output is characterized by an abrupt jump at the signal onset, a log-like compression during the static portion of the signal (auditory "adaptation"), and a sharp drop and recovery at the signal offset. This design has a remarkable cabability to accentuate signal change, such as onset and offsets, and compress stationary regions such as steady portions of vowels, in a manner a static design could not achieve. That is, this capability cannot be obtained with a simple static nonlinearity, such as a fixed logarithmic operation, that is more standardly used in auditory modeling.

Figure 7: Response (b) of the DPK auditory model to a 875 Hz sinewave (a) centered on the 9th bandpass filter.
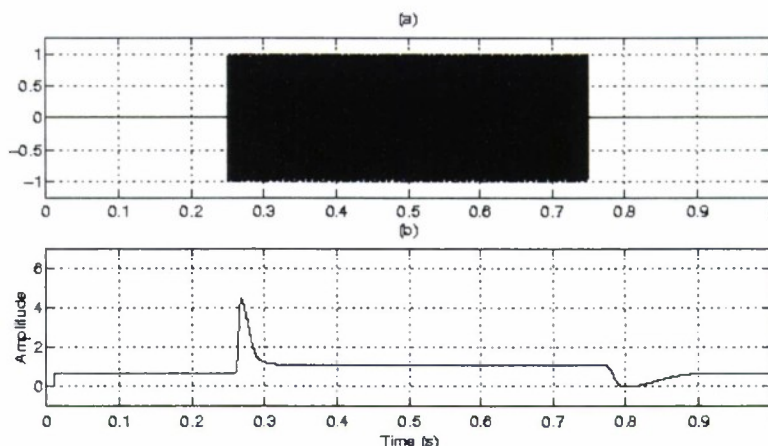
On real speech, as in the synthetic case, while the static regions are found to be close to log-compressed, rapidly-varying regions are accentuated, as illustrated in the example of Figure 8. In the figure, the output envelope of the 2nd (near 200 Hz) and the 11th (near 2000 Hz) are shown for both the full DPK auditory model and the DPK model with the adaptive nonlinearity replaced by a static log-compression. The later case acts similar to the computation of the standard mel-scale filter energies. Our first observation is that the presence of the adaptive nonlinearity clearly accentuates the speech dynamics, relative to the log-compression. Onsets and transient events are accentuated. Furthermore, different filters reveal different events; in the example, the filter #2 output envelope shows onsets (dashed lines) while filter #11 better reveals transient events (dotted lines), in particular the plosives. Another interesting observation is that we see in low-frequency filter outputs, semblances of the response to a single sine wave, as depicted in the box superimposed on the #2 filter-output envelope. This is because the low-frequency filters are narrow enough to pass a few, and sometimes a single, sine wave.
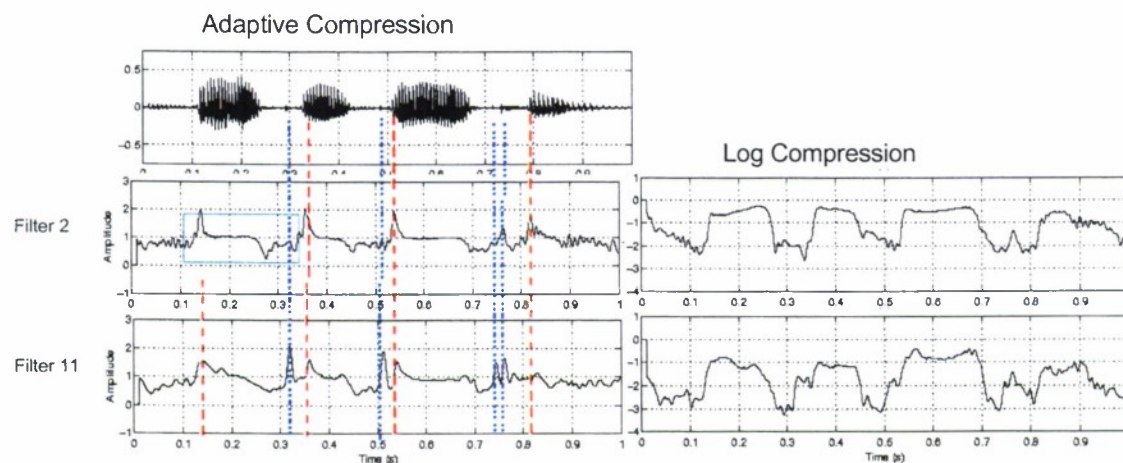

Figure 8: Illustration of band-pass filter-output envelope of the DPK auditory model with the adaptive nonlinearity (left bottom panels) and with static log-compression (right panels) . Upper left panel shows the original speech waveform. Two filter outputs are depicted.

3.0 RECOGNITION WITH FULL FRONT-END AUDITORY MODEL

Speaker recognition experiments were performed using the Lincoln GMM-UBM recognition system [Reynolds, 1995]. The training and testing utterances are taken from a subset of the 1998 NIST evaluation corpus; both correspond to electret handsets, but possibly different telephone numbers. 50 target speakers are used for each gender with 262 test utterances for males and 363 for females. Target training utterances are 2 min in duration and test utterances 30 s in duration. Background models are trained using the 1996 NIST evaluation corpus and 1996 NIST evaluation development corpus. The experiments were performed using both standard cepstral mean subtraction and RASTA to account for telephone channel degradation. Energy measures from the mel-filters and the DPK auditory front-end filters were downsampled and transformed to a 19-element feature vector, allied with delta-cepstra, through the discrete-cosine transform. We performed recognition comparisons at a 5-ms analysis frame interval because this frame interval gives a slight performance gain over the standard 10-ms interval for the auditory features and comparable performance with a 10-ms interval for mel-cepstral features. Observe that the standard RASTA filter (designed for a 10-ms frame) requires a re-design. This was performed by an explicit factor-of-two upsampling of the standard RASTA filter impulse response to create an unconventional FIR RASTA filter.

The result of a speaker recognition experiment with features derived from the DPK auditory model is shown in Figure 9 where we see Detection Error Rate (DET) performance for male speakers, relative to the recognition with the standard mel-cepstrum. Although there is a small relative decrease in performance with the use of the DPK features over some parts of the DET curve, the important observation is that the new auditory features provide a good basis for speaker recognition. An important question is whether these new features provide new, "orthogonal", information that the mel-cepstrum does not provide. Toward this end, we fused the scores (with equal weight) obtained from the two feature sets, with the result of a reasonable performance gain over either feature set alone, as shown in Figure 9. There appears then to be some orthogonality between the two feature sets.

Figure 9: Illustration of speaker verification improvement via score fusion of DPK auditory and mel-cepstrum features. Fusion of scores provides a performance improvement over either feature set alone.
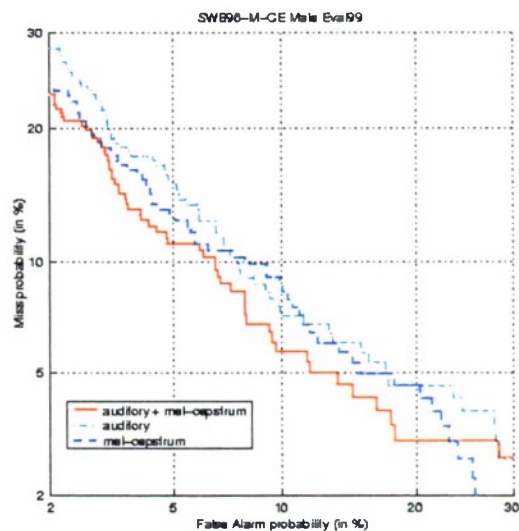
It should be noted that in recognition with the new auditory-based features, we have applied both Cepstral Mean Subtraction (CMS) and RASTA. Application of RASTA, however, is in some sense self-defeating because RASTA low-pass-filters temporal trajectories, thus perhaps reducing the rapid changes we are attempting to exploit in recognition. In future work, removing RASTA may therefore further improve "orthogonality" with respect to the standard mel-cepstrum. Finally, we emphasize that another caveat in the above experiments is that, through an "exact inverse" post-filter (Section 2.1), we have forced the temporal resolution of the baseline auditory filtering to be equivalent to that of the mel-cepstrum. We enforced this restriction in part to provide a known (preliminary) starting point and in part to avoid the need for multi-resolution sampling of filter-bank outputs.

4.0 A HIGH-LEVEL AUDITORY CHANNEL EXTENSION

Although we have seen some performance gain with the use of new auditory features for speaker recognition, we have barely exploited the temporal dynamics of the complete chain of aural processing by the human; that is, the richness of the temporal trajectories was lost in a simple envelope sampling. Indeed, Dau, Kollmeier, and Kohlrausch [Dau, Kollmeier, and Kohlrausch, 1997] realized this limitation in seeking to extend the DPK model and, as a result, proposed an interesting extension of the original model. This extension was motivated by the physiological studies of Schreiner and Langner [Schreiner and Langner, 1988], showing that amplitude modulations in the cochlea channels are passed up the auditory pathway to a higher-level region of the brain referred to as the *inferior colliculus* (IC). These amplitude modulations are represented in a systematic way *orthogonally to the tonotopical organization of the basilar membrane*. The experiments were performed on a cat's auditory system which is known to be similar to that of the human. The placement along the auditory pathway of the inferior colliculus relative to that of the cochlear front end is illustrated in Figure 10.
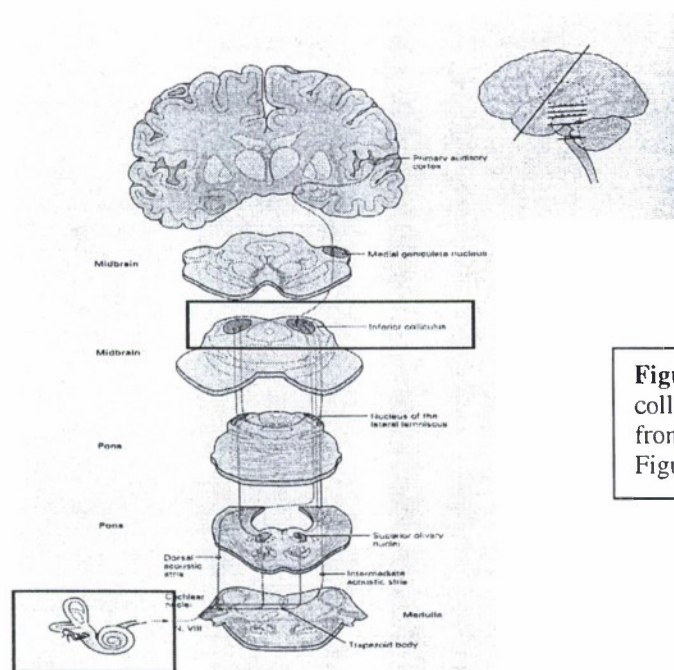

Figure 10: Location of the inferior colliculus relative to that of the cochlear front end along the auditory pathway. Figure from [Hudspeth, 2000].

9

4.1 Model of the inferior colliculus (IC)

To further clarify the importance the inferior colliculus (IC), in the Schreiner and Langner work, the IC of the cat revealed a linear array of cells that corresponds to the tonotopic organization of the cochlea filters of the basilar membrane. By "tonotopic" we mean the spatial dimension of the basilar membrane being related monotonically to frequency. Moreover, each IC cell is tuned to different temporal modulation frequencies of the envelope of a cochlea filter-output signal. The stimulus in these experiments consisted of sine waves with carriers equal to the characteristic frequency of the IC cells. Each sine wave was amplitude-modulated with the frequency of the amplitude modulation selected as the experimental variable, and the response of each cell was measured with the changing frequency of the amplitude modulation. Modulation transfer functions were obtained, i.e. ,the response as a function of modulation frequency, and best modulation frequencies were then established for each IC cell. Best modulation frequencies were found to scan the range [90, 1000] Hz, with a concentration between 30 and 300 Hz. The shape of the modulation transfer functions is primarily band-pass, although high-pass, low-pass, and band-reject were also observed. Because of some *redundancy* in the modulation transfer functions, best modulation frequencies may appear in more than one IC cell, covering a range of different bandwidths.

For many IC cells, the best modulation frequency was characterized by a strong synchronization to the signal envelope, similar to a phase synchrony at the lower cochlea signal level [Pickles, 1988]. (Another property of the IC, perhaps related to this synchrony, is the frequency occurrence of "intrinsic oscillations" that are speculated to provide a delay line for correlation analysis within the IC region.) In many cases, this synchrony appears strong just after the stimulus onset for modulation frequencies above the best modulation frequency and wane thereafter. Synchrony also deteriorates with increasing modulation frequency. On the average, the mean and upper best modulation frequency for a channel increased with the characteristic frequency of the IC cell. Although the best modulation frequency is concentrated in the range [30, 300] Hz, as mentioned, the best frequency was found to go as high as 1000 Hz, thus providing high temporal resolution. This temporal envelope representation of the inferior colliculus thus complements the more well-established spectral-based processing at the lower-level cochlea. One approach to realize this, perhaps, "orthogonal" information for recognition is to perform a frequency analysis of each temporal trajectory as through, for example, a filter-bank analysis [Dau, Kollmeier, and Kohlrausch, 1997]. Features can then be obtained from this filter-bank output.

In invoking the desired frequency analysis of each filter-bank-output envelope, one preliminary "nested" analysis configuration that we are proposing is illustrated in Figure 11 where a 12-element filter bank over a 1000-Hz range (Figure 12) is allied to the envelope of each filter output. "Nested" refers to the output of each nonlinear auditory channel (denoted as *nac* in Figure 11) being processed by a second-stage filter bank. The bandpass filters that we implemented are similar to those used by Dau, Kollmeier, and Kohlrausch [Dau, Kollmeier, and Kohlrausch, 1997], being 1st-order recursive with different time constants corresponding to different filter bandwidths. Each filter-bank output is complex and can be represented by a complex phasor: $a[n]\exp(j\theta[n])$ with magnitude $a[n]$ and phase $\theta[n]$. Above 10 Hz, we sample only the magnitude, $a[n]$, of each channel at a 5-ms frame rate. Below 10 Hz, we sample the quadrature representation, $a[n]\cos(j\theta[n])$, because the magnitude function at such low frequency is said to not be well estimated. In addition, this preserves signal phase in the low-frequency region and thus is consistent with physiological evidence of such phase preservation [Pickles, 1988]. In this sampling strategy, *no attempt has been made to address aliasing due to*

undersampling. The extraction of features from the $24 \times 12 = 288$ element array of temporal envelopes is clearly unwieldy. Nevertheless, such a 2-D array is likely the information-(modulation) rich signal-network structure that is used by the higher (brain) levels of the auditory system.
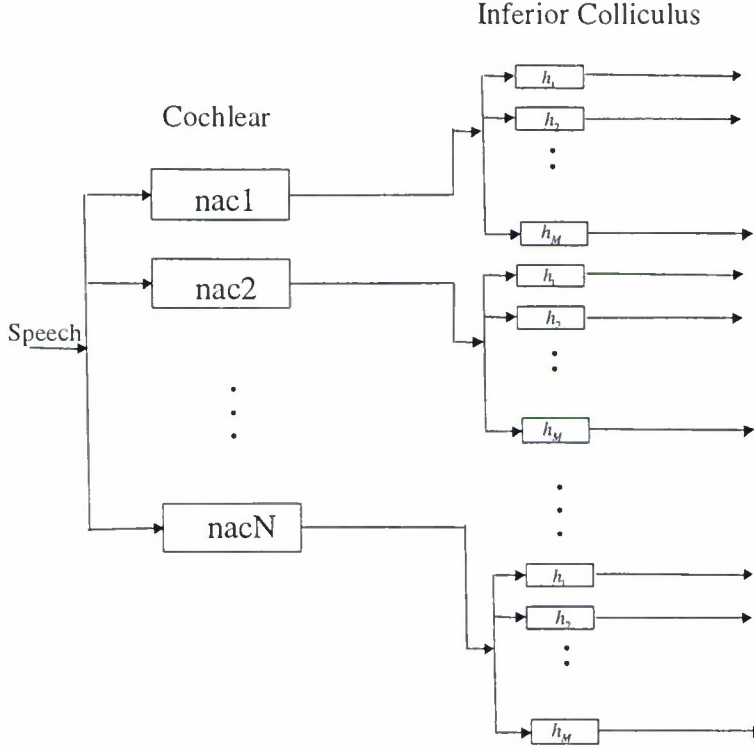
Figure 11: A nested-channel extension of the DPK front-end auditory model. The outputs of each front-end nonlinear auditory channel (nac1, nac2 ...nacN), i.e., the processor of Figure 3, are further analyzed by the filter bank $h_1, h_2 ... h_M$ (Figure 12). "Nested" refers to the output of each nac being processed by a second-stage filter bank.
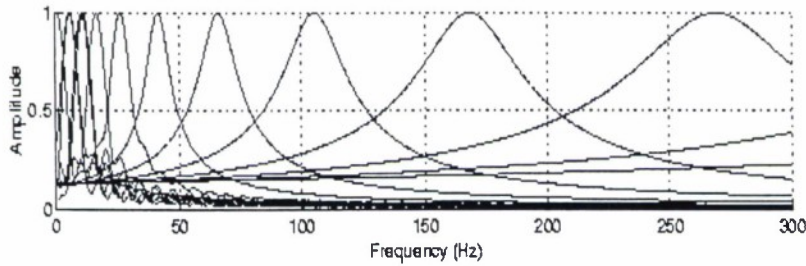
Figure 12: The band-pass filters used to measure common channel modulation (in the nested configuration of Figure 11) are uniformly spaced by 5 Hz with a 5-Hz bandwidth up to 10 Hz and have logarithmic spacing and bandwidth above 10 Hz up to 1000 Hz. To preserve phase in the low-frequency region, the quadrature version of the filter output is sampled up to 10 Hz, while the filter output envelope is sampled above 10 Hz. No attempt has been made to address aliasing due to under-sampling. Only the first 300 Hz of the filter bank is shown.

11

4.2 Feature Extraction

One approach that we are proposing to reduce dimensionality is seen in the summation post-network in Figure 13. The summation implements an average of the filtered trajectory outputs, corresponding to the same modulation frequency across channels. Interesting properties of this configuration is that it is consistent with the above observed *redundancy* across IC filters, *enhances common modulation across frequency,* and may provide a reasonable reduction of the information that is orthogonal to the tonotopical organization of the basilar membrane (e.g., information provided by the mel-cepstrum). It is interesting to observe that our extraction of common modulation is also consistent with recent theories of auditory scene analysis [Bregman, 1990].
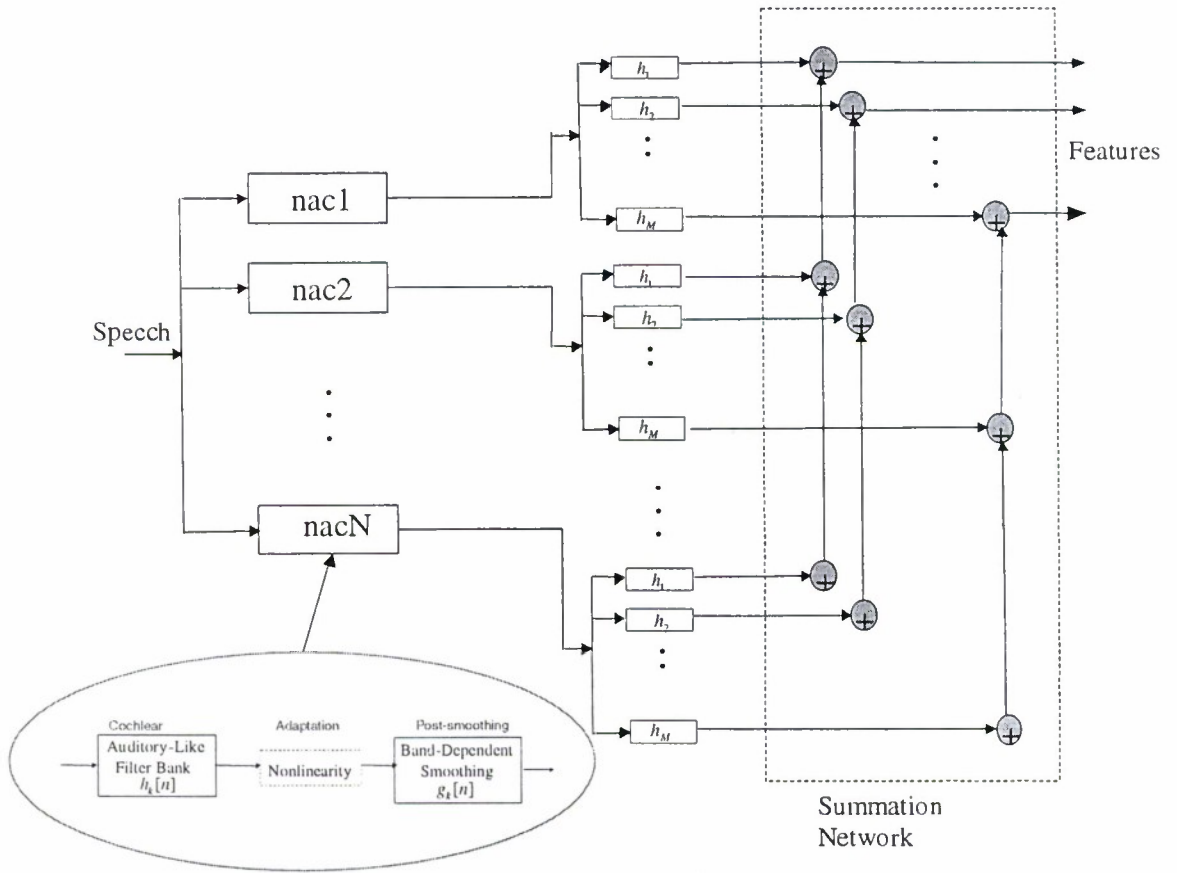

Figure 13: Reduction of dimensionality in the nested–channel extension of the DPK auditory model. For each front-end center frequency, the output envelopes of each filter $h_1, h_2, ... h_M$ are summed to enhance common modulation and to yield a reduced feature set.

12

An example of the nested auditory output for speech input is shown in Figure 14, consisting of a particularly noisy Switchboard (NIST 98) example. In the figure we see two sets of summed filter outputs: from band-pass filter #4 at about 25 Hz and from band-pass filter #12 at about 800 Hz. We have turned off the post-smoothing in the auditory front end to maintain signal dynamics prior to summation. We see that the summation accentuates dynamics common across cochlear-filter outputs and is robust in the presence of noise. The modulation structure for the high-frequency summation reflects fine transient, transition, and pitch structure, while the modulation structure for the low-frequency summation reflects a broader look at common modulation across frequency. Observe also that the low-pass filter reveals a sine-like response as depicted in the boxes of Figure 14. Clearly, the current fixed 5-ms frame sampling imparts aliasing due to under-sampling of the modulation from high-frequency summation.

Figure 14: Summation output (c) of one low-frequency filter (filter #4) and (b) of one high-frequency filter (filter #12) of the high-level inferior colliculus-model filter bank (Figure 12). The original noisy Switchboard passage is given in panel (a). As an aside, it is interesting to observe in panel (c) the two responses enclosed in the boxes are similar to that of Figure 7b where a single sine wave enters a narrowband filter.
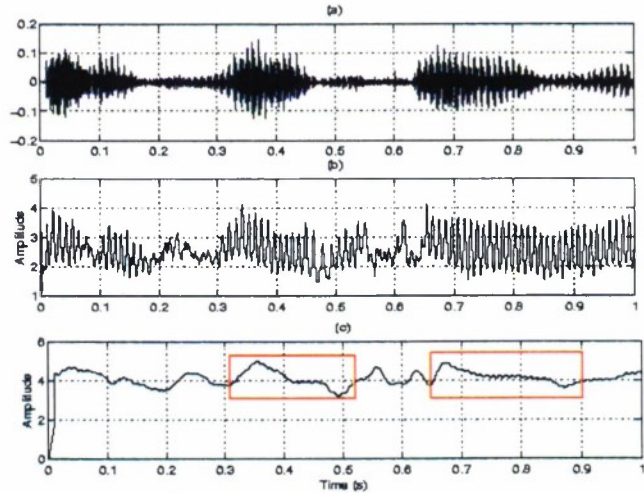
4.3 Speaker Recognition

The result of automatic speaker recognition with the inferior colliculus-based features is shown in Figure 15 where we see performance for male speakers, relative to recognition with the DPK front-end auditory-model and the mel-cepstrum features. Although there is relatively low performance with the inferior colliculus-based features, fusion of scores from these features with the two other feature sets provides a good incremental performance gain in speaker recognition, relative to the solid curve of Figure 9. Consistent with the physiological observations of Schreiner and Langner [Schreiner and Langner, 1988], there is additional "orthogonality" relative to the two previous spectrally-based feature sets.

Figure 15: Illustration of speaker verification improvement via fusion of scores from the DPK auditory front end (0.25 weight), the high-level inferior colliculus model (0.25 weight), and the mel-cepstrum (0.50 weight). Fusion of scores provides a useful performance improvement over the mel-cepstrum alone.
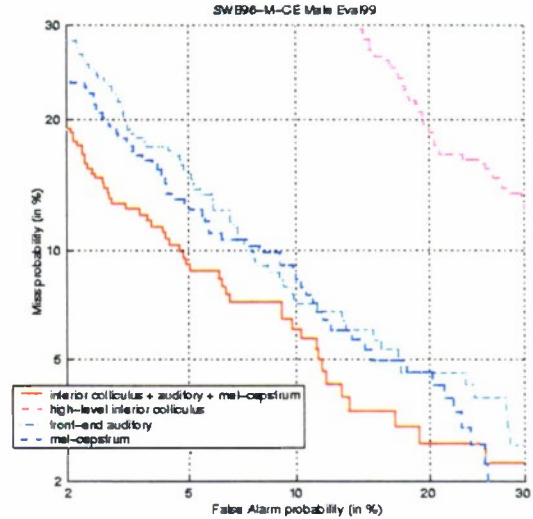
5.0 FUTURE DIRECTIONS

We began this research effort with the objective of extracting speech source and vocal tract parameters based on new linear and non-linear speech production models, and to apply these techniques to improve speaker recognition in degrading environments characterized by noise and distortion. The auditory model-based approach was introduced because of its known ability to represent temporal structure on both a fine and coarse scale, as introduced by complex nonlinear production models, and because of its recognition robustness [Tchorz and Kollmeier, 1999], [Schmidt-Nielsen and Crystal, 2000]. We felt this perception-based, non-parametric approach is a more solid foundation than the parametric approach that originally motivated this study in which secondary pulse times provided temporal feature streams [Quatieri, Jankowski, and Reynolds, 1994] because the later approach requires explicit formant, pitch, and voicing estimation that is not desirable under degrading conditions.

In summary, from preliminary speaker recognition results, we have shown some degree of orthogonality with the mel-cepstrum of the two new auditory-based feature sets: (1) The nonlinearity of the auditory front-end provides accentuated dynamics not seen by a conventional auditory front-end model and (2) A frequency analysis and integration of common modulation of the auditory channels of (1) further emphasizes temporal structure. Fusion of likelihood scores from these two feature sets with mel-cepstral scores provides an encouraging recognition performance gain over use of the conventional mel-cepstrum alone.

Our approach provides an important new direction in speech feature representation with many possibilities and questions remaining. Here we name several:

1. The particular auditory dynamics for the nonlinear auditory component (Figures 3 and 6) were selected to mimic the work of Tchorz and Kollmeier [Tchorz and Kollmeier, 1999] who are developing features for speech recognition. The auditory system, on the other hand, is known to sometimes adapt itself to the task at hand, and thus it behooves us to optimize the nonlinear dynamics for speaker recognition.

2. The postfilters used in our auditory model (Figures 3 and 5) were selected to preserve a constant temporal resolution across frequency. Clearly, this imposes an unnecessary limitation. Likewise (as

14

alluded to at the end of Section 3), an additional limitation has been invoked in the use of RASTA with features derived from the nonlinear auditory model.

3. Our summation method of feature reduction from the 288 channels of the high-level auditory model of Figures 11 and 13 are preliminary and heuristic. One can conceive of a multitude of other approaches. One alternate approach we are currently exploring relates sub-groups of channels to auditory precepts (which may relate to abstract articulatory precepts). Each sub-group will form a feature stream, and likelihood scores from these feature streams will be fused.

4. In general, although improving resolution, our new approach is still lacking in a clear understanding and implementation of time- and frequency-resolution and sampling. With respect to resolution, we seek to include hierarchical temporal information, going beyond simple delta cepstra that "look" over about 50 ms with sub-groups [of (3) above] that integrate temporal information over longer time durations. With respect to sampling, our current frame-based, uniform sampling needs to be generalized to a non-frame-based, non-uniform approach, as motivated by the multi-resolution of Figure 14.

5. Exploiting a large training corpus, such as the NIST 2001 extended corpus (e.g., 45 min training sessions), for representing temporal structure on different scales. Evidence of the importance of larger time scales for speaker recognition has been demonstrated in idiolectal speech representations which requires phone or speech recognition [Doddington, 2001], [Andrews, Kohler, Campbell, 2001]. Our multi-level auditory-based approach provides the possibility of capturing idiolectal differences without the need of speech or phone recognition.

6. Testing the new auditory-based features in degrading environments, relative to the mel-cepstrum, including channel mismatch conditions, where the channels for the training and test data are mismatched. Additional speaker identification experiments will be carried out on the MIST (Multilingual Interoperability of Speech Technology) data being collected by the NATO IST-TG01 group, as this data becomes available.

7. Investigating the effect of channel normalization methods such as znorm, hnorm, and tnorm.

8. Revisit early work on temporal features derived from the distributed-source paradigm described in the introduction to this report. This would involve developing a nonlinear aeroacoustic speech production model accounting for the spatially distributed sound sources and the associated multi-source excitation induced by non-acoustic fluid motion in the vocal tract, building upon the previous work of Sinder, Krane, and Flanagan [Sinder, Krane, and Flanagan, 1998], [Sinder, 1999]. In general form, this model will include n sources S1, S2,...,Sn distributed along the vocal tract, each contributing to the speech production via corresponding transfer functions H1, H2, ..., Hn. A first experiment will be specializing this general nonlinear distributed source model to the case of two sources, and using this model to develop robust techniques for estimating the effects of secondary pulse excitations, as were initially shown in the earlier work at Lincoln Laboratory [Quatieri, Jankowski, and Reynolds, 1994]. Extensions to more than two sources, and relation to our above modulation auditory modeling, would then follow. For the speaker identification experiments, the distributed source model techniques will be combined with other robust speaker identification techniques.

REFERENCES

[Andrews, Kohler, Campbell, 2001] W. Andrews, M. Kohler, J. Campbell, "Phonetic speaker recognition," *Proc. Eurospeech 2001*, Aalborg, Denmark, September 3-7, 2001.

[Doddington, 2001] G. Doddington, "Some experiments on idiolectal differences among speakers", http://www.nist.gov/speech/tests/spk/2001/doc/, March 1, 2001.

[Bregman, 1990] A.S. Bregman, "Auditory Scene Analysis", MIT Press, Cambridge, MA, 1990.

[Dau, Puschel, and Kohlrausch, 1997] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the "effective" signal processing in the auditory system", *J. Acoust. Soc. of America*, vol. 99, no. 6, June 1996.

[Dau, Kollmeier, and Kohlrausch, 1997] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation (Parts I and II)",, *J. Acoust. Soc. of America*, vol. 102, no. 5, June 1997.

[Hudspeth, 2000] A.J. Hudspeth, "Hearing", Chapter in Principles of Neural Science, 4th Edition, Edited by E.R. Kandel, J.H. Schwartz, and T.M. Jessell, McGraw-Hill, New York, N.Y., 2000.

[Maragos, Kaiser, and Quatieri, 1996] P. Maragos, J. Kaiser, T.F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis ", *IEEE Trans. Signal Processing*, Vol. 41, No. 10, October 1993, pp. 3024-3051.

[Pickles, 1988] A. Pickles, *An Introduction to Auditory Physiology*, Academic Press, 2nd Edition, New York, NY, 1988.

[Quatieri, Jankowski, and Reynolds, 1994] T.F. Quatieri, C.R. Jankowski, and D.A. Reynolds, "Energy onset times for speaker identification", *IEEE Signal Processing Letters*, vol. 1, no. 11, pp. 160-162, 1994.

[Reynolds, 1995] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication*, vol. 17, pp. 91-108, August 1995.

[Schmidt-Nielsen and Crystal, 2000] A. Schmidt-Nielsen and T. Crystal, "Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data", *Digital Signal Processing*, Special Issue: NIST 1999 Speaker Recognition Workshop, Editors: J. Campbell and J. Schroeder, Academic Press, vol. 10, no. 1-3, pp. 249-266, January/April/July 2000.

[Schreiner and Langner, 1988] C. Schreiner and G. Langner, "Periodicity coding in the inferior colliculus of th cat: 1. Neuronal Mechanism", *J. Neurophysiol.*", vol. 60, pp. 1799-1822, 1988.

[Sinder, 1999] D.J. Sinder, *Speech Synthesis Using an Aeroacoustic Fricative Model*, Ph.D. Thesis, Rutgers University, New Brunswick, NJ, October 1999.

[Sinder, Krane, and Flanagan, 1998] D.J. Sinder, M.H. Krane, and J.L. Flanagan, "Synthesis of fricative sounds using an aeroacoustic noise generation model", Proceedings of the International Congress on Acoustics and the Acoustical Society of America (ICA/ASA), Seattle, Washington, June 1998.

[Strobe and Alwan, 1998] B. Strobe and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition", *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 5, Sept. 1997.

[Tchorz and Kollmeier, 1999] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition", *J. Acoust. Soc. of America*, vol. 106, no. 4, October 1999.

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 17 May 2002	3. REPORT TYPE AND DATES COVERED Final Report

4. TITLE AND SUBTITLE

Nonlinear Auditory Modeling as a Basis for Speaker Recognition

5. FUNDING NUMBERS

C—F19628-00-C-0002

6. AUTHOR(S)

T.F. Quatieri

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
244 Wood Street
Lexington, MA 02420-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

Final Report

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory (AFRL/IFEC)
Rome Research Site
Bldg. 240, 32 Brooks Road
Rome, NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESC-TR-2001-065

11. SUPPLEMENTARY NOTES

None

AD #
A 402327

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

In this report, we develop a front-end nonlinear auditory model based on recent work of Dau, Puschel, and Kohlrausch (DPK) [Dau, Puschel, and Kohlrausch, 1997]. An important aspect of the model is the robust accentuation of temporal change in a signal at the cochlea level that forms the basis of a feature set for automatic speaker recognition. Preliminary speaker recognition experiments with the DPK front-end auditory model give performance close to that from the standard mel-cepstrum. Fusion of scores from the mel-cepstrum and the DPK front-end auditory model, however, is shown to give a useful performance gain relative to the standard mel-cepstrum alone. The dynamics provided by the nonlinear auditory model, therefore, appears to provide some "orthogonality" to that of the more static mel-cepstral representation.

In addition, in this report, we provide initial development of new "common modulation" features based on modeling a more central region of auditory processing in the brain's *inferior colliculus* than the low-level auditory front-end. These higher-level features rely on the DPK auditory model as a foundation for further analysis of low-level temporal trajectories. This new feature representation is an important research direction and provides additional feature "orthogonality" to front-end auditory processing, as exhibited in improved speaker recognition performance with fusion of scores from low-level and high-level feature sets.

14. SUBJECT TERMS

15. NUMBER OF PAGES
26

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Same as Report